Digital Object Identifier 10.1109/ACCESS.2017.DOI

Appendix A - Comparison of methodologies

Table A1: Comparison of methodologies [1, 2, 3, 4, 5]

Method	Utility	Limitations	Potential applicability to study	Need satisfied
Bibliometrics	• Provides measures of growth in a given subject domain (i.e. different technologies) based on patent counts • Citation analysis enables clustering based on communities of authors, publication years, key references, or citation pathways	• Limited by statistical data available from publications (gaps may be present) • Patent citations susceptible to noise (i.e. topics focusing on one specific technical aspect of a patent, but unrelated to the main concept being discussed) • Human bias present in selection of patents to cite • Publication data is viewed separate from historical events (i.e. requires additional qualitative data to establish context & links to global trends/influences). Does not automatically identify generalisable patterns without this additional context	• Extraction of growth curves for deriving relationships to substitution patterns • Definition of historical TLC measures from extracted patent datasets (i.e. 10 patent indicators)	Ability to extract appropriate data measures for use in classification
Pattern recognition (clustering)	• Preferred technique for classification: less susceptible , compared to other techniques, to outliers &model changes as data points added • Better suited to complex interactions between variables • Allows labelling when training data unavailable	• Cross-validation and ranking required to identify best combination of TLC measures to use to ensure most robust out-of-sample performance • Information loss associated with outliers • No. of clusters may need to be specified (risk of human error) • Does not consider phase variance directly	• Measurement of similarity between life cycle metrics for different technologies & substitution modes • Cross-validation tests: assessment of classification scheme validity & identification of most robust class predictors	Ability to determine most appropriate predictors (i.e. data measures) to use in classification model
Pattern recognition (functional linear regression)	• Takes phase variance between dissimilar time series into account • Produces time series as output (required for subsequent causal model building)	• Regression not typically regarded as good for classification models: every time a new point is added, model has to be updated • Cross-validation and ranking required to identify best combination of TLC measures to use to ensure most robust out-of-sample performance • More prone to errors & outliers than clustering	• Creation of a predictive life cycle classification model, based on real-world timings	Ability to build time-dependent classification model
Multi-level regression	• Accounts for variation arising from groupings • Provides statistically efficient estimates of regression coefficients • Provides correct standard errors, confidence intervals, and significance tests (typically more 'conservative' than those obtained when ignoring clustering) • Enables model variation to be more accurately mapped to specific influences • Enables a relative ranking of available metrics (i.e. individual patent indicators)	• Requires strong assumptions about the structuring of data (i.e. clearly bounded hierarchies) • More computationally expensive	• Creation of a predictive life cycle classification model, based on real-world timings & most robust predictors	Ability to build time-dependent classification model, based on most robust predictors, enabling extrapolation of TLC trends from preliminary data when phase variance is considered
Classical statistical regression (non-hierarchical)	• Many well understood statistical tests available for determining standard errors, confidence intervals, & significance tests	• Does not account for variation between groups, which may have a strong effect on predicted trends (i.e. risk of invalid results) • No automatic means of extending predictions to new groups • Estimates of varying effects can be noisy, especially with few observations per group • Risk of over-fitting without some form of cross-validation • Regression not typically regarded as good for classification models: every time a new point is added, model has to be updated • More prone to errors & outliers than clustering • Does not consider phase variance	• Creation of a predictive life cycle classification model (without considering phase variance)	Ability to extrapolate TLC trends from preliminary data, when phase variance is not an issue ***(Not applicable here)***

References

[1] H. Goldstein, Multilevel statistical models. John Wiley & Sons, 2011, vol. 922.

[2] A. Gelman and J. Hill, Data analysis using regression and multilevel/hierarchical models. Cambridge university press, 2006.

[3] WIPO, “WIPO Guide to Using Patent Information,” Geneva: Word Intellectual Property Organization, 2010.

[4] H.-Y. Yeh, Y.-S. Sung, H.-W. Yang, W.-C. Tsai, and D.-Z. Chen, “The bibliographic coupling approach to filter the cited and uncited patent citations: a case of electric vehicle technology,” Scientometrics, vol. 94, no. 1, pp. 75–93, 2013.

[5] P. Joshi, “When to Use Linear Regression, Clustering, or Decision Trees,” 2017, accessed: June 2018. [Online]. Available: https://dzone.com/articles/decision-trees-vs-clustering-algorithms-vs-linear